DOCUMENT RESUME

ED 407 414 TM 026 435

AUTHOR Gray, B. Thomas
TITLE Controversies regarding the Nature of Score Validity: Still Crazy after All These Years.
PUB DATE 23 Jan 97
NOTE 21p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Testing; *Evaluation Methods; Reliability; *Scores; *Test Interpretation; *Validity
IDENTIFIERS *Controversy

ABSTRACT

Validity is a critically important issue with far-reaching implications for testing. The history of conceptualizations of validity over the past 50 years is reviewed, and 3 important areas of controversy are examined. First, the question of whether the three traditionally recognized types of validity should be integrated as a unitary entity of construct validity is examined. Second, the issue of the role of consequences in assessing test validity is discussed, and finally the concept that validity is a property of test scores and their interpretations, and not of tests themselves is reviewed. The shift from the "trinitarian" doctrine of content, construct, and criterion validity has meant that the distinctions between different types of validity have been replaced by recognition of the varieties of evidence required in the validation process. It is universally acknowledged that validity is a crucial consideration in evaluating tests and test applications. It is also generally stated that a true validation argument is an unending process. Exploring new ideas about the nature of validity itself is just a part of this process. (Contains 50 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 407 414

Running Head: CONTROVERSIES REGARDING SCORE VALIDITY

Controversies Regarding the Nature of Score Validity: Still Crazy After All These Years

B. Thomas Gray

Texas A&M University 77843-4225

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
B. THOMAS GRAY
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 23, 1997.

BEST COPY AVAILABLE

TM026435

Abstract

Validity is a critically important issue with far-reaching implications for testing. The history of conceptualizations of validity over the past 50 years is reviewed, and three important areas of controversy are examined. First, the question of whether the three traditionally recognized types of validity should be integrated as a unitary entity of construct validity is examined. Second, the issue of the role of consequences in assessing test validity is discussed, and finally, the concept that validity is a property of test scores and their interpretations, and not of tests themselves is reviewed.

Controversies Regarding the Nature of Score Validity:
Still Crazy After All These Years

The most recent edition of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement and Education, 1985) included the bold statement that: "Validity is the most important consideration in test evaluation" (p. 9). It seems likely that the same point will be reiterated, perhaps verbatim, in the forthcoming revised edition of the same work. The importance indicated by such a strong declaration is reinforced by the fact that no new test can be introduced without a manual that includes a section on validity studies, and no text on testing and/or psychometrics is considered complete without at least one chapter addressing the topic of validity.

In 1949, Cronbach (p. 48) stated that the definition of validity as "the extent to which a test measures what it purports to measure" was commonly accepted, although he preferred a slight modification: "A test is valid to the degree that we know what it measures or predicts" (p. 48). Cureton (1951) provided similar commentary: "The essential question of test validity is how well a test does the job it is employed to do... Validity is therefore defined in terms of the correlation between the actual test scores and the 'true' criterion scores" (pp. 621, 623). The enduring definition given by Anastasi (cf., 1954, p. 120; Anastasi & Urbani, 1997, p. 113)--"Validity is what the test measures and how well it does so"--is cited quite widely.

It is interesting to note that Cronbach, one of the most prominent voices in the field of psychometrics, and a widely respected authority on the topic of validity, has of late tended to avoid the problem of defining the term after the 1949 statement cited above (cf., 1988, 1989). In

1971 (p. 443), however, he provided an insightful statement that foreshadowed some of the controversy of the future: "Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score."

Exceptions can be found to the apparent conservatism seen in the definitions cited above. Perhaps most notable is Messick (1989a, p. 13), who stated that, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." This reflects much of the debate and controversy to be found in the literature of the past several years, indicative perhaps of greater intellectual movement in the field than would be implied by the previous paragraph.

It is certainly beyond the scope of the present paper to present a comprehensive review of the topic of validity. The purpose instead is to focus on a few more obvious points of controversy. Three areas in particular will be addressed: (a) the status of the different "types" of validity; (b) the issue of what is sometimes referred to as "consequential validity"; and (c) the persistence of illogical statements taking the broad form of, "The test is valid."

The above discussion illustrates the considerable differences in the way validity is conceptualized by different authorities. Some have changed their views little over the past 40+ years, while others have been advocating markedly different views, a few for many years. Since the roots of many of the shifts in thinking which are occurring today can be found in earlier works, a brief historical review of the ways in which the validity formulations have developed is helpful in understanding both the current controversies and the persistent themes. For further detail, the interested reader is referred particularly to the extensive discussions of the topic which

have appeared in the three volumes of Educational Measurement (Cronbach, 1971b; Cureton, 1951; Messick, 1989) published thus far.

Conceptualizations of Validity: An Historical Sketch

By the early part of the 1950's a plethora of different "types" of validity (factorial, intrinsic, empirical, logical, and many others) had been named (see Anastasi, 1954). Among those whose contributions continue to be acknowledged are Gullickson (1950), Guilford (1946), Jenkins (1946), and Rulon (1946). Typical formulations recognized two basic categories, which Cronbach (1949) termed "logical" and "empirical" forms of validity. The former was a rather loosely organized, broadly defined set of approaches, including content analyses, and examination of operational issues and test taking processes. Test makers were expected to "make a careful study of the test itself, to determine what test scores mean" (Cronbach, 1949, p. 48). Much of what has since become known as content validity is found with in this broad category.

Empirical validity placed emphasis on the use of factor analysis (e.g., Guilford's 1946 factorial validity), and especially on correlation(s) between test scores and a criterion measure (Anastasi, 1950). Cureton (1951) devoted several pages to various issues concerning the criterion, and the influence of this approach is seen in its widespread use even today (Cronbach, 1989), despite some apparent limitations. For example, Cureton's assertion quoted above is easily (although perhaps slightly outlandishly) refuted by noting that a positive correlation could be obtained between children's raw scores on an achievement test and their heights. This is not to say that correlational studies are useless, but rather that their indiscriminate application can sometimes yield uninteresting results.

Several interesting and important political developments converged to influence events

relating to validity conceptualization (Benjamin, 1996; Cronbach, 1988, 1989). In the late 1940's, the academically-oriented APA, was attempting to draw the membership of the new Association for Applied Psychology back into its ranks. The two groups combined into what was supposed to be a more broadly-oriented APA, and work was begun on establishing an appropriate code of ethics addressing both scientific and applied concerns. A committee was established in 1950 to develop standards for adequate psychological tests, and their work was published in 1954. At that time four categories of validity were defined: content, predictive, concurrent, and construct.

That basic outline is still in use today (AERA et al., 1985), essentially unchanged, with the exception that in 1966, the revised edition of the Standards combined the predictive and concurrent validity categories into the single grouping called criterion validity. The 1954 standards, which were actually referred to as "Technical Recommendations" in their first edition, were quickly followed up by the publication in 1955 of Cronbach and Meehl's landmark paper, "Construct Validity in Psychological Tests" (see Thompson & Daniel, 1996). The construct of "Construct Validity" was elucidated more clearly, including introduction of the concept of the "nomological net." The latter is described as the interrelated "laws" supporting a given construct. Cronbach (1989) later presented this in somewhat less strict terms, acknowledging the impossibility with most constructs used in the social sciences of attaining the levels of "proof" demanded in the harder sciences.

Soon thereafter, Campbell (1957) introduced into the validation process the notion of falsification , and discussed the importance of testing "plausible rival hypotheses." This was explained in further detail by Campbell and Fiske (1959) in their important paper (Thompson & Daniel, 1996) introducing the multitrait-multimethod approach and the notions of convergent and

divergent (or discriminant) validity. There have been objections to some of the applications of this technique, particularly insofar as it can and sometimes does become a rather rote exercise, which will therefore produce only vapid results. Nonetheless, the multitrait-multimethod approach, like correlational studies, enjoys widespread appeal nearly 40 years after its introduction.

Construct Validity as a Unifying Theme

The so-called "trinitarian doctrine," which conceptualizes validity in three parts, has been a fundamental part of the Standards since their inception (or, to be picky, at least since 1966). This doctrine is therefore presented as standard fare in most textbooks which cover the topic of validity. Anastasi, for example, has followed the same outline in her widely-used textbook Psychological Testing since 1961, despite her commentary (1986; Anastasi & Urbani, 1997; also see below) that there is considerable overlap between these different categories, and that the distinctions are sometimes confusing. Interestingly, Cronbach also continued to follow the same outline in the various editions of his very popular text, Essentials of Psychological Testing (1960, 1971a, 1984), despite his strong protestations on its inadequacy noted below.

While Guion (1980) is often cited as the first to have made the suggestion that the three "types" should be unified under the single heading of construct validity, in fact there have been others who have argued the "unitary" position for many years. This has been based not only on the difficulty in distinguishing between three types, but more importantly, following careful consideration of the relevant concepts. Among the more powerful voices have been those of Messick (1965, 1975, 1989a, 1989b, 1995) and Cronbach (1971b, 1988, 1989). The groundswell was apparently sufficient to have prompted a subtle but important shift between the 1966 and

1985 editions of the Standards in that the latter used terms such as content-related and criterion-related validity, rather than simply content and criterion validity. By implication, in 1985 these were being considered as different parts of the single construct of validity, rather than as distinct and separate entities. One would hope that the committee charged with the currently revision of the Standards will continue this trend.

Messick (1995, p. 7) argued that content validity "does not qualify as validity at all," because it provides evidence based solely on judgment evidence (i.e., expert opinion) regarding the relevance of the test material and the representativeness of its content to the domain of interest. This is not evidence to support any sort of inferences or conclusions based on test scores (Messick, 1975, 1995). The argument could be advanced that domain relevance and representativeness are necessary bases for any construct. However, Cronbach (1988, 1989) has pointed out that content is assessed only with respect to the construct in which we are interested. Thus, with any but the most simple assessment of content the focus will necessarily shift beyond content to a consideration of the construct itself.

Cronbach (1988, 1989) went on to present the compelling argument that assessment of a criterion measure also necessarily implies that it is being assessed with respect to some sort of construct, since this certainly does not occur in a vacuum. The construct is central to any validity argument since it is the interpretations based on the test scores, not the scores themselves, that are of primary concern; the interpretations are almost by definition construct-based. Messick (1989b, 1994) further contended that limiting a validation argument to a criterion-based study is too narrow, and the generalizability of the argument will therefore be limited. This is because such a study involves only "selected parts of the test's external structure," and "there are as many

criterion-related validities for the test scores as there are criterion measures and settings" (Messick, 1989b, p. 7).

This is not to say, however, that examination of criterion-related information is of no value. Indeed, like reliability, validity is a function of scores (and more specifically, score interpretation), and is not a characteristic of a test; that is to say, validity coefficients will vary from sample to sample, from population to population, from occasion to occasion. Thus, examination of criterion-related validity(-ies), and especially selection of an appropriate criterion measure where possible, will dictate the level of generalizability appropriate for a given test.

Moss (1992, 1995) has outlined several different schemes for subdividing the concept of validity. Various authors have argued, sometimes strongly (e.g., Shepard, 1993), in favor of discarding the traditional three types, but there appears to be little agreement concerning an alternative system. For instance, Messick (1989a, 1989b) provided two different systems that were published in the same year. As Moss (1995) pointed out, it is questionable whether a new classification scheme would be able to overcome the momentum of more than 40 years of usage of a "trinitarian" system. The most critical point, however, is that the categories should be seen as various kinds of information pertaining to a unitary notion of validity.

Values and Consequent Validity

A number of authors have discussed the nature of validity interpretations and their relationship to factors beyond simply content- and/or criterion-based data. For example, Messick (1975, 1980, 1989a, 1989b, 1994, 1995) has long pointed out that validity decisions, by their nature, are value laden in that they involve judgments based on evidence. The decision as to how large a correlation coefficient must be in order to be considered important is an example, as is the

interpretation of factor pattern and structure coefficients, both involve a series of value-based judgments. Values influence the way questions are framed, and also the way decisions "based on the results" are made. (The use of quotation marks in the last sentence is perhaps an overly cynical way of pointing out the oft-noted fact that, despite the ideal approach of validation through the standard scientific approach of testing plausible rival hypotheses, researchers typically are not particularly good at doing this with their own work).

Cronbach (1988) explicitly outlined a series of considerations, which he termed "perspectives," that strongly affect validity decisions; these include influences from the political, economic, and legal arenas, and are often value-based. Arguing from this viewpoint, he, Messick (1975, 1989a, 1989b, 1995) and others (e.g., Kane, 1992; Moss, 1992, 1994, 1995; Shepard, 1993) have concluded not only that such values must be considered in validity discussions, but that the consequences of the tests in question must also be assessed.

Messick (1975, 1989b, 1995) has perhaps most fully developed this line of thinking which emphasizes the importance of test consequences, and has summarized his ideas in a four-celled table that has appeared in many of his publications beginning in 1965. Under his scheme, test application is divided into two categories, interpretation and use; the other axis of the table is divided according to evidential and consequential bases for validity. Each of the four cells is termed a facet of validity. Construct validity is the evidential basis for test interpretation, and construct validity plus value implications are the consequential basis. Construct validity plus relevance and/or utility make up the evidential basis for test use, while all three together with social consequences represent the consequential basis.

This formulation was first presented more than 30 years ago (Messick, 1965), and has

been thoroughly analyzed and discussed. Nonetheless, a few questions of simple logic arise. In particular, given that value implications may be found in many if not most aspects of validity arguments, particularly those that involve judgment, it seems likely that judgments concerning the relevance/utility of a test will necessarily have value implications. It is therefore not entirely clear what will separate the consequential basis for test interpretation from the evidential or even the consequential basis for test use. Shepard (1993) agreed with Messick's basic contention concerning the importance of considering the consequences of test application, although she disagreed with the way he has gone about it. Her primary objection was that dividing validity in this fashion detracts from the view of validity as a single, unitary entity, which to her is of preeminent importance. Messick himself seems somewhat unclear on this point, in that in one paper (1989b) he specifically stated that assessment of consequences should not be taken to represent a separate type (facet, perspectives, or whatever) of validity, yet later in the same paper he used the term "consequential validity" more than once.

It is perhaps important to note that many of the stronger arguments favoring the notion of assessing consequences come from scholars primary concerned with issues of program evaluation. Brandon, Lindberg, and Wang (1993) presented an example of integrating student feedback and other considerations regarding potential consequences in their development of a new curriculum for their medical school. They claimed good success with this approach, although it is far from clear how it might generalize to situations involving single individuals. In either case, several questions arise (see Lees-Haley, 1996). For any given situation, the worker will inevitably be faced with the problem of deciding which consequences are of greatest concern, that is to say, one would have to determine the consequences to whom, as judged by whom, and over what length of

time (i.e., short term and/or long term).

Two hypothetical examples might serve to clarify some of the potential dilemmas. In redesigning a medical school curriculum (with apologies to Brandon et al., 1993), the parties most directly involved, the faculty and the students, might ultimately decide that anatomy should be eliminated from the required training. After all, anatomy is a rather dull topic, dominated by rote memorization, and is generally unpleasant to learn; for similar reasons, it is also generally unpleasant to teach. The short term consequences to both faculty and students of the standard anatomy requirement are apparently unfavorable. On the other hand, this writer would most definitely be less than eager to be treated by a physician who graduated from a medical school that did not require its students to learn anatomy.

On the level of the individual, consider someone being administered a series of tests to determine if s/he should be admitted to a state mental institution. If the decision is "yes," then the consequences to that person, both short term and probably also long term, are unfavorable. At the same time, however, the consequences to his or her family, to any other social agencies and institutions with whom s/he comes in contact, and even to the community in which s/he lives, are also relevant, and may be quite severe should the decision be, "no."

Few would argue that concerns such as these are unimportant. The question that must be asked, though, is whether or not such considerations belong under the heading of "Validity." Maguirre, Hattie, and Haig (1994) suggest that consequences are not appropriate grist for this mill, and that validity issues should be limited to questions more directly pertinent to measurement. Indeed, it would seem that as debate grows concerning the various aspects of validity, measurement issues are perhaps becoming a "red-headed step-child," relegated to the

corner to be left to its own devices (Zimiles, 1996).

The Test is NOT Valid

It has already been well established that validity statements apply only to the population on which the particular study was based, and it is therefore patently illogical to make statements indicating that a particular test is valid. This has recently been explicitly stated by several authors (e.g., Thompson, 1994a, 1994b; Wainer & Braun, 1989), and in fact, statements to this effect may be found in earlier works, such as Cronbach's (1971b) paper. There are even hints that Cureton had reached similar conclusions as early as 1951:

> The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third. Hence, we cannot label the validity of a test as "high" or "moderate" or "low" except for some particular purpose. (p. 621)

In fact, there may be rare instances in which it is justifiable to conclude that a given test is valid for a given range of purposes. Cronbach (1988) briefly alluded to this when he wrote of the extensive work of Schmidt, Hunter, and their colleagues on the generalization of general ability tests for employment purposes (Hunter & Schmidt, 1982; Sackett, Tenopyr, Schmitt, & Kahn, 1985; Schmidt, Pearlman, Hunter, & Hirsh, 1985). However, the amount of the work involved in developing Cronbach and Meehl's (1955) "nomological net" to support such a claim must not be overlooked. The resources necessary to conduct sufficient research to be able to establish this net will certainly not be possible in most cases.

Summary and Discussion

It is interesting to see how some ideas have changed over the years, and how others, considered new and innovative, can actually be traced back several decades. Rogers (1996)

presented a brief review of the history of standards for evaluating programs, and identified eight trends pertaining to validity which have occurred over the past 50 years. These are paraphrased here:

1. a shift from focusing on the test to the testing process;

2. the inclusion of various modes of measurement beyond just pencil and paper tests;

3. the inclusion of logical forms of argument in validity questions;

4. a shift to from trinitarian to unitarian doctrine (with distinctions between different types of validity replace by recognition of the varieties of evidence required in validation process);

5. the inclusion of evidence regarding context effects (relevant to generalizability) in test assessment;

6. introduction of an emphasis on standards for producers and also for users of tests.;

7. explicit recognition of the role of values in validity; and

8. recognition of the need to consider utility and social consequences of test use.

This paper has focused particularly on items 4, 7, and 8.

Cronbach (1989, p. 147) described the "sad fact that almost every psychologist writing about [construct validity] applies to it the word 'confusing,'" and the same commentary applies to the traditional doctrine that identified three types of validity: content, criterion, and construct (cf., Anastasi, 1954, 1988). Today most authorities have pretty well agreed that such an approach is difficult to support, "that content and criterion validities are no more than strands within a cable of validity argument" (Cronbach, 1988, p. 4), and that it is more parsimonious to consider validity as a single, unitary concept. Various systems have been proposed to subdivide validity into different facets, aspects, or perspectives, although general agreement is most definitely lacking on

this point. The general consensus nonetheless is that, however it may (or may not) be divided, the different parts represent lines of evidence pointing toward the single construct.

It has also been fairly well demonstrated that, contrary to prevailing opinion of 40 to 50 years ago, no mode of scientific inquiry is devoid of the influence of values. From this recognition, several authors have argued that one must include consequences of the application of a given test as an aspect of the validity of that application. This is a much more controversial area, for which there is far less consensus. It would seem that, at a minimum, many portions of this argument must be clarified before consequential validity is universally accepted as a facet of validity that must always be considered.

Finally, the illogic of the mantra, "The test is valid" was discussed. That statements of such form persist despite the strong reasons for not using them is testimony to the inertia that accrues to any long-standing practice. The phenomenon is similar to the persistence of what Cronbach (1989, pp. 162-163) termed "empirical miscellany" and "unfocused empiricism" seen most clearly in the accumulation of various correlation coefficients that serves as the "validity argument" in many (perhaps most) test manuals.

The controversies that persist are welcomed. Consider, for example, that the basic outline of validity presented by Anastasi did not change for over 30 years (cf. Anastasi, 1961, 1988; Anastasi & Urbani, 1997). This is strongly suggestive of stagnation in thinking, a condition which is only alleviated by the challenge of new ideas. Not all of the new ideas discussed in the works reviewed in the present paper are necessarily useful. At least most of those that are not useful will not survive the tests of time.

It is universally acknowledged that validity is a crucial consideration in evaluating tests

and test applications. It is also generally stated that a true validation argument, rather than resulting from a single study, such as might be found in a newly published test manual, is an unending process. Contending with new ideas regarding the nature of validity itself is just a part of this process.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement and Education (1985). Standards for educational and psychological testing. Washington, DC: Author.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51, 201-238.

American Psychological Association (1966). Standards for educational and psychological tests and manuals. Washington, D.C.: Author.

Anastasi, A. (1954). Psychological testing. New York: Macmillan.

Anastasi, A. (1961). Psychological testing (2nd ed.). New York: Macmillan.

Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Anastasi, A. (1986). Evolving concepts of test validation. Annual Review of Psychology, 37, 1-15.

Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). New York: Macmillan.

Benjamin, L. T. (1996). The founding of the American Psychologist: The professional journal that wasn't. American Psychologist, 51, 8-12.

Brandon, P. R., Lindberg, M. A., & Wang, Z. (1993). Involving program beneficiaries in the early stages of evaluation: Issues of consequential validity and influence. Educational Evaluation and Policy Analysis, 15, 420-428.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity in the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cronbach, L. J. (1949). Essentials of psychological testing. New York: Harper & Row.

Cronbach, L. J. (1960). Essentials of psychological testing (2nd ed.). New York: Harper & Row.

Cronbach, L. J. (1971a). Essentials of psychological testing (3th ed.). New York: Harper & Row.

Cronbach, L. J. (1971b). Test validation. In R. L. Thorndike (Ed.),. Educational measurement (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1984). Essentials of psychological testing (4th ed.). New York: Harper & Row.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence. Measurement theory and public policy (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1954). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cureton, E. F. (1951). Validity. In E. F. Lindquist (Ed.), Educational measurement (1st ed., pp. 621-694). Washington, DC: American Council on Education.

Guilford, J. P. (1946). New standards for test evaluation. Educational and Psychological Measurement, 6, 427-439.

Gulliksen, H. (1950). Instrinsic validity. American Psychologist, 5, 511-517.

Guion, R. M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-398.

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. D. Dunnette & E. A. Fleishman (Eds.), Human capability assessment. Hillsdale, NJ: Lawrence Erlbaum.

Jenkins, J. G. (1946). Validity for what? Journal of Consulting Psychology, 10, 93-98.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112, 527-535.

Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. American Psychologist, 51, 981-983.

Maguire, T., Hattie, J., & Haig, B. (1994). Alberta Journal of Educational Research, 40, 109-126.

Messick, S. (1965). Personality measurement and the ethics of assessment. American Psychologist, 20, 136-142.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50, 741-749.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62, 229-258.

Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

Moss, P. A. (1995). Themes and variations in validity theory. Educational Measurement: Issues and Practice, 14(2), 5-12.

Rogers, W. T. (1996). The treatment of measurement issues in the revised Program Evaluation Standards. Journal of Experimental Education, 63(1), 13-28.

Rulon, P. J. (1946). On the validity of educational tests. Harvard Educational Review, 16, 290-296.

Sackett, P. R., Tenopyr, M. L., Schmitt, N., & Kahn, J. (1985). Commentary on forty questions about validity generalization and meta-analysis. Personnel Psychology, 38, 697-798.

Schmidt, F. L., Pearlman, K., Hunter, J. E., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. Personnel Psychology, 38, 697-798.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of research in education (Vol. 19, pp. 405-450). Washington, DC: American Educational

Research Association.

Thompson, B. (1994a, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1994b). Guidelines for authors. Educational and Psychological Measurement, 54(4), 837-847.

Thompson, B., & Daniel, L. G. (1996). Seminal readings on reliability and validity: A "hit parade" bibliography. Educational and Psychological Measurement, 56, 741-745.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. Phi Delta Kappan, 75, 200-214.

Zimiles, H. (1996). Rethinking the validity of psychological assessment. American Psychologist, 51, 980-981.